Московский государственный технический университет им. Н. Э. Баумана

Курс «Технологии машинного обучения»

Отчёт по рубежному контролю №1

«Технологии разведочного анализа и обработки данных.»

Вариант № 5

Выполнила:
Буйдина К.А.
группа ИУ5-63Б

Проверил:
Гапанюк Ю.Е.

Дата: 11.04.25

Подпись:

Дата:

Подпись:

Москва, 2025 г.

Задание:

Номер варианта: **5**

Номер задачи: **1**

Номер набора данных, указанного в задаче: **5**
https://www.kaggle.com/mohansacharya/graduate-admissions (файл Admission_Predict.csv)

Для студентов групп ИУ5-63Б, ИУ5Ц-83Б - для произвольной колонки данных построить график "Ящик с усами (boxplot)".

Задача №1.

Для заданного набора данных проведите корреляционный анализ. В случае наличия пропусков в данных удалите строки или колонки, содержащие пропуски. Сделайте выводы о возможности построения моделей машинного обучения и о возможном вкладе признаков в модель.

Ход выполнения:

Задача 1

Для заданного набора данных проведите корреляционный анализ.

В случае наличия пропусков в данных удалите строки или колонки, содержащие пропуски.

Сделайте выводы о возможности построения моделей машинного обучения и о возможном вкладе признаков в модель.

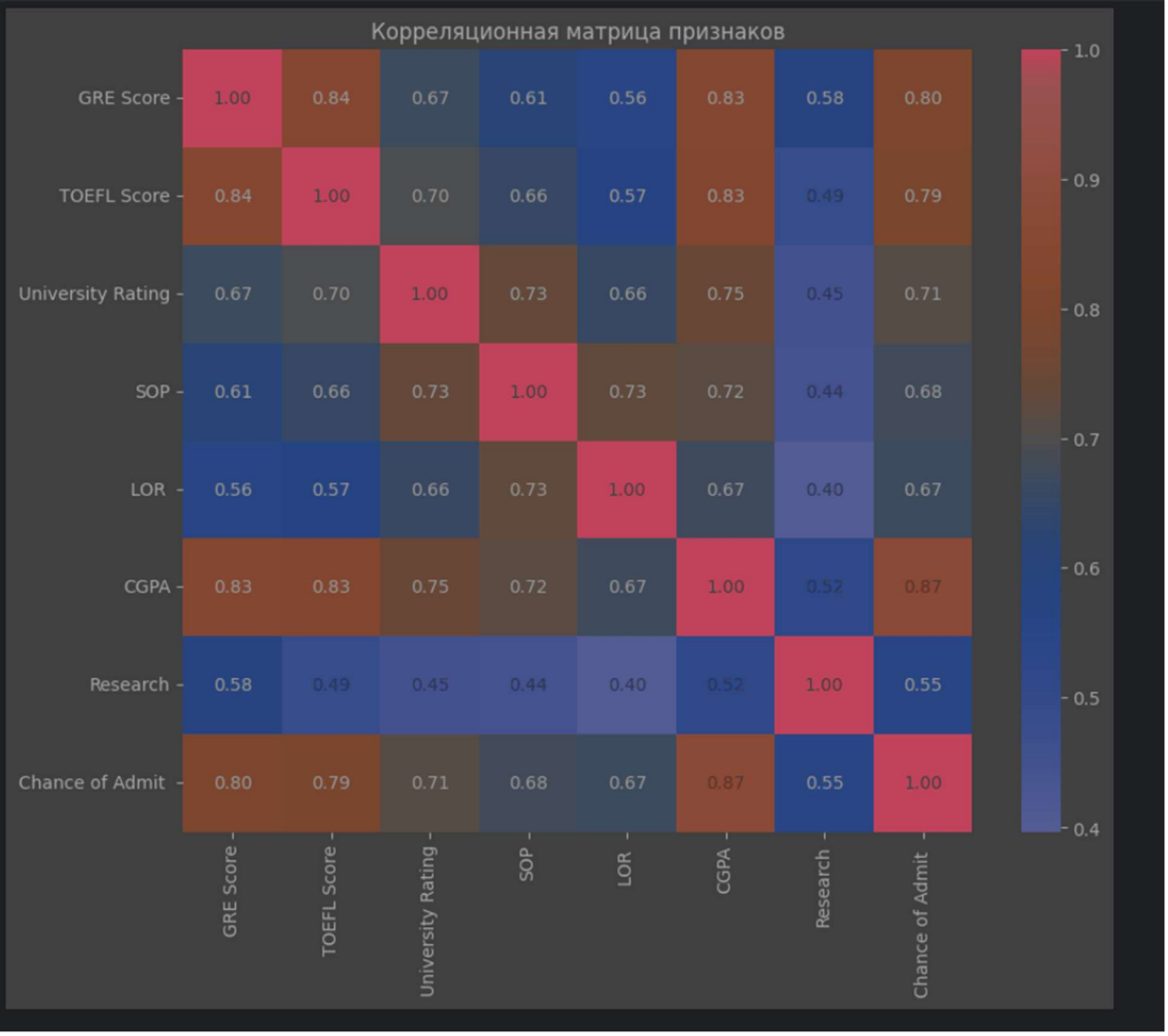
```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

df = pd.read_csv('Admission_Predict.csv')
df.drop('Serial No.', axis=1, inplace=True)
print("Количество пропущенных значений в каждой колонке:")
print(df.isnull().sum())
```

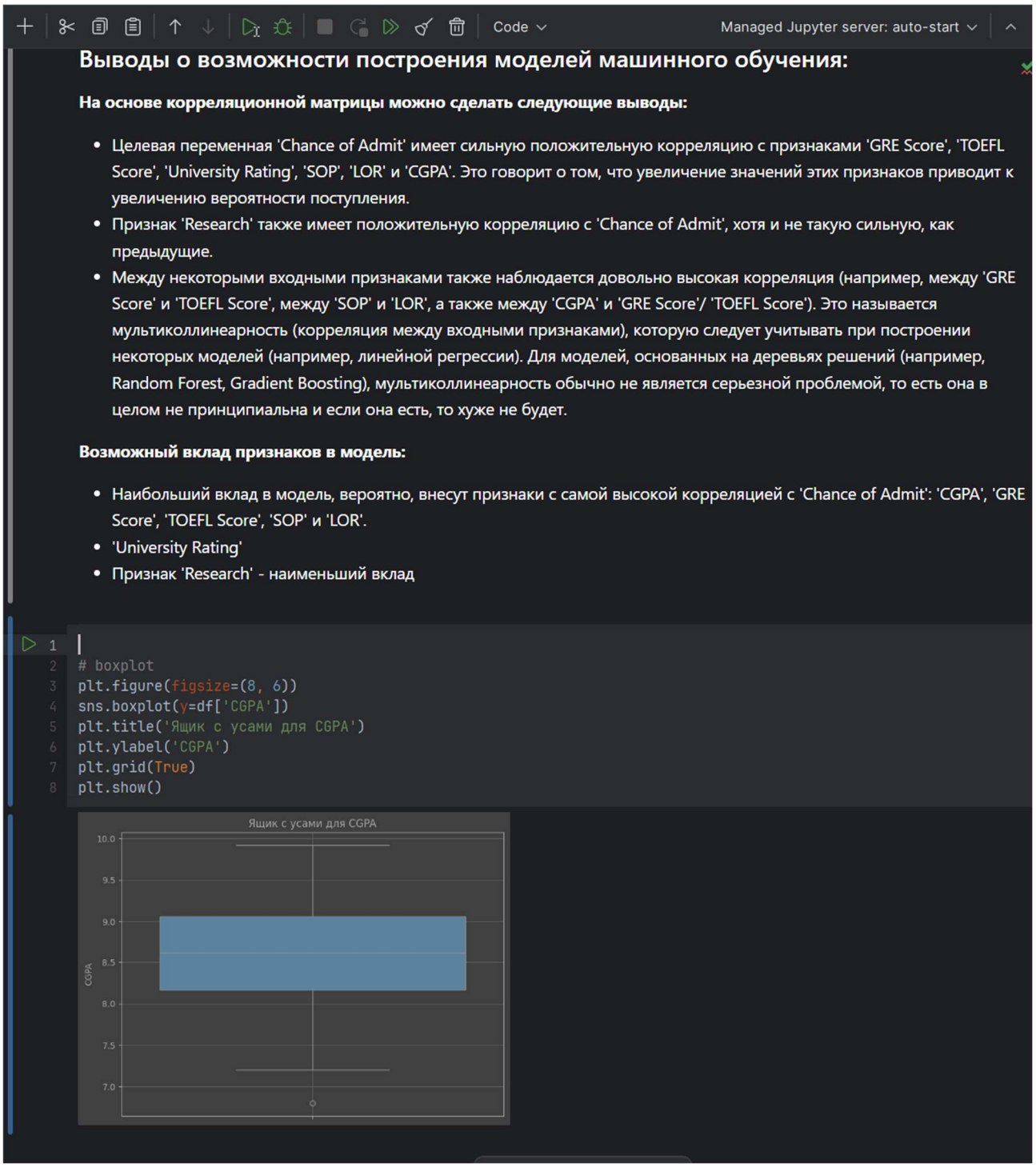
```
Количество пропущенных значений в каждой колонке:
GRE Score            0
TOEFL Score          0
University Rating    0
SOP                  0
LOR                  0
CGPA                 0
Research             0
Chance of Admit      0
dtype: int64
```

вывод - пропусков не было

```python

# корреляция
correlation_matrix = df.corr()
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Корреляционная матрица признаков')
plt.show()
```

```python

# корреляция
correlation_matrix = df.corr()
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Корреляционная матрица признаков')
plt.show()

```




Code
Managed Jupyter server: auto-start
Выводы о возможности построения моделей машинного обучения:
На основе корреляционной матрицы можно сделать следующие выводы:
Целевая переменная 'Chance of Admit' имеет сильную положительную корреляцию с признаками 'GRE Score', 'TOEFL Score', 'University Rating', 'SOP', 'LOR' и 'CGPA'. Это говорит о том, что увеличение значений этих признаков приводит к увеличению вероятности поступления.
Признак 'Research' также имеет положительную корреляцию с 'Chance of Admit', хотя и не такую сильную, как предыдущие.
Между некоторыми входными признаками также наблюдается довольно высокая корреляция (например, между 'GRE Score' и 'TOEFL Score', между 'SOP' и 'LOR', а также между 'CGPA' и 'GRE Score'/ 'TOEFL Score'). Это называется мультиколлинеарность (корреляция между входными признаками), которую следует учитывать при построении некоторых моделей (например, линейной регрессии). Для моделей, основанных на деревьях решений (например, Random Forest, Gradient Boosting), мультиколлинеарность обычно не является серьезной проблемой, то есть она в целом не принципиальна и если она есть, то хуже не будет.
Возможный вклад признаков в модель:
Наибольший вклад в модель, вероятно, внесут признаки с самой высокой корреляцией с 'Chance of Admit': 'CGPA', 'GRE Score', 'TOEFL Score', 'SOP' и 'LOR'.
'University Rating'
Признак 'Research' - наименьший вклад
1
# boxplot
plt.figure(figsize=(8, 6))
sns.boxplot(y=df['CGPA'])
plt.title('Ящик с усами для CGPA')
plt.ylabel('CGPA')
plt.grid(True)
plt.show()
Ящик с усами для CGPA
CGPA
10.0
9.5
9.0
8.5
8.0
7.5
7.0